E

Adjustment of Observed Intake Data to Estimate the Distribution of Usual Intakes in a Group

An individual's actual intake varies considerably from one day to the next, but it is usual or long-term average intakes that are of interest in assessing and planning dietary intakes to ensure nutrient adequacy for individuals or groups. As explained in a previous report (IOM, 2000a), serious error in the assessment of nutrient inadequacy or excess can occur if the dietary intake data examined do not reflect usual intakes. This poses a major obstacle to the assessment of an individual's nutrient intake because his or her usual intake is generally poorly estimated from only a few days of observation, yet more extensive data collection is rarely feasible. Assessments of nutrient adequacy among groups are facilitated by the availability of statistical adjustment procedures to estimate the distribution of usual intakes from observed intakes, as long as more than one day of intake data has been collected for at least a representative subsample of the group. These procedures do not yield estimates of usual intake for particular individuals in the group, but the adjusted distribution of intakes is appropriate for use in analyses of the prevalence of inadequate or excess intakes in the group.

In recent years a number of different statistical procedures have been developed to estimate the distribution of usual intakes from repeated short-term measurements (Hoffmann et al., 2002). Two commonly used adjustment procedures are described here: the National Research Council (NRC) method and the Iowa State University (ISU) method. Both procedures are based on a common conceptual foundation, but the ISU method includes a number of statistical enhancements that make it more appropriate for use with large population surveys. The NRC method is simpler and may be more appropriate than the ISU method for use with small samples (those with less than 40 to 50 individuals). However, neither method is without limitations.

THE NATIONAL RESEARCH COUNCIL METHOD

Conceptual Underpinnings

In assessing nutrient adequacy it is necessary to estimate usual intake. However, usual intake cannot be inferred from measures of observed intake without error. For any one individual,

Observed intake = usual intake + measurement error

The observed variance ($V_{observed}$) of a distribution of intakes for a group based on one or more days of intake data per individual is the sum of the variance in true usual intakes of the individuals who comprise the group (e.g., the between-person or interindividual variance, $V_{between}$) and the error in the measurement of individuals' true usual intakes. Error arises both because of the normal variation in individuals' intakes from one day to the next and because of random error in the measurement of intake on any one day. It is referred to as the within-person, day-to-day, or intraindividual variance (V_{within}) (NRC, 1986).

$$V_{observed} = V_{between} + V_{within} + V_{underreporting}$$

The observed distribution of intakes will be wider and flatter than the true distribution of usual intakes as a result of the presence of within-person variance. However, assuming that the within-person variation is random in nature, the estimate of mean intake for the group will not be influenced by this variance.

If multiple days of intake data per individual are averaged, and the distribution of intakes in the group is constructed from the means of each individual's multiple intakes, then the error variance (e.g., within-person variance) diminishes as a function of the number of days of intake data per person. Thus, as the number of days of data per person increases, the distribution of observed intakes (expressed as the individuals' observed mean intakes over the days of data collection) becomes a better and better approximation of the true distribution of usual intakes in the group.

The NRC method (NRC, 1986) is typically applied to a data set comprising multiple days of intake data for a sample of individuals, ideally with an equal number of observations per individual. This method of estimating the distribution of usual intakes works by first partitioning the observed variance into its between- and within-person components, and then shifting each point in the observed distribution closer to the mean by a function of the ratio of the square roots of the between-person variance ($V_{between}$) and observed variance ($V_{observed}$). In this way, the method attempts to remove the effect of within-person variation on the observed distribution. The variance of the adjusted distribution should represent $V_{between}$.

Application

The steps in the NRC method are outlined below. The method is illustrated using data on the zinc intakes of 46 women recorded over three, nonconsecutive, 24-hour dietary intake recalls (a subsample of women drawn from a earlier study by Tarasuk and Beaton [1999]).

Step 1. Examine normality of distribution and transform data if necessary.

This adjustment procedure depends on the properties of a normal distribution, yet the observed distribution of intakes for most nutrients is likely to be positively skewed. This is because the distribution is naturally truncated at 0 (i.e., reported intakes cannot fall below this value) but has no limit at the upper end. Thus it is imperative that the normality of the 1-day intake data be assessed. (This can be accomplished through the NORMAL option in PROC UNIVARIATE in SAS.) If departures from normality are detected, the data should be transformed to approximate a normal distribution. The most appropriate transformation will depend on the shape of the original distribution; it may have a logarithm, square root, or cubed root relationship.

Note that for this example, the assessment of normality is conducted on all 138 days of recall data (e.g., 46 women multiplied by 3 days). The Shapiro-Wilk statistic, *W*, provides one measure of the normality of the data (Tarasuk and Beaton, 1999). For the raw data, $W = 0.85$ (versus a value of 1 for normally distributed data), and the distribution departs significantly from normality ($p < 0.0001$). A visual inspection of the plotted data reveals that they are right-skewed. Through a process of trial and error, a more normal distribution is achieved by applying a cubed root transformation to these data.

The W of the transformed data is 0.99 (p = 0.1812). The next two steps in this adjustment procedure are conducted using the transformed data.

Step 2. Estimate the within- and between-person variance.

Some statistical packages have procedures for partitioning the variance of the observed data into the within- and between-person variance components (e.g., PROC VARCOMP in SAS). This can also be easily accomplished using the analysis of variance procedures available in most statistical packages by conducting a simple one-way ANOVA with subject ID included as a categorical or class variable. A sample program for SAS is presented at the end of this appendix. When the raw data are transformed to better resemble a normal distribution, this step is conducted on the transformed data.

Two values are extracted from the ANOVA output. The mean square error or unexplained variance (e.g., the variance in the observed daily intakes that is not accounted for by between-subject differences) represents the within-subject variance in the 1-day data. The mean square model (e.g., the mean square associated with the subject ID variable entered into the ANOVA) represents the observed variance of the 1-day data. Because the adjustment procedure is applied to an individual subject's mean intakes over the period of observation, both the mean square model and mean square error need to be divided by the mean number of days of intake data per subject to obtain the $V_{observed}$ and V_{within} for this distribution (e.g., $V_{observed}$ = mean square model/n and V_{within} = mean square error/n). $V_{between}$ can be estimated by subtracting V_{within} from $V_{observed}$, as follows:

$$V_{between} = (\text{mean square model} - \text{mean square error})/n$$

where n is the mean number of days of intake data per subject in the sample. $V_{between}$ represents the "true" variance of the distribution of usual intakes. Each of these variance estimates can be expressed as a standard deviation by simply taking the square root of the variance.

Table E-1 presents the output for the ANOVA procedure as applied to this example. The mean number of days of intake data per subject is three. In this example, $V_{observed}$ = 0.24633584/3, V_{within} = 0.13375542/3 and $V_{between}$ = (0.24633584 – 0.13375542)/3.

TABLE E-1 ANOVA of Zinc Intake of 46 Adult Women, Shown for Data Transformed Using Cubed Roots

Source	Degrees of Freedom	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Model	45	11.08511265	0.24633584	1.84	< 0.0069
Error	92	12.30549834	0.13375542		
Corrected total	137	23.39061099			

Step 3. Adjust individual subjects' mean intakes to estimate the distribution of usual intakes.

Each subject's mean intake is now adjusted by applying the following formula:

$$\text{Adjusted intake} = [(\text{subject's mean} - \text{group mean}) \times (SD_{between}/SD_{observed})] + \text{group mean}$$

where $SD_{between}$ is the square root of $V_{between}$ and $SD_{observed}$ is the square root of $V_{observed}$. This equation effectively moves each point in the distribution of observed intakes closer to the group mean, but it does not shift the group mean. If the distribution of 1-day data was transformed prior to partitioning the variance (Step 2), the equation is applied to the individual subject and group means calculated from the transformed data (Step 3), and the resultant distribution needs to be transformed back prior to use (see Step 4). If the data were not transformed, however, the adjusted intakes calculated from this equation now represent the estimated distribution of usual intakes.

Step 4. If the original data have been transformed, transform the adjusted intake back to the original units.

If the original data were transformed in order to satisfy the necessary assumption of normality, the adjusted data need to be transformed back into the original units prior to their use for nutrient assessment. Back-transforming refers to the application of the inverse function of the original transformation. In this example, the original data were transformed using cubed roots; the back transformation raises subject's adjusted intakes to the power of three. The process of transforming data, adjusting it, and then back-transforming it is

TABLE E-2 Observed Distribution of 3-day Mean Zinc Intakes (mg) and Estimated (Adjusted) Distribution of Usual Intakes for a Sample of 46 Women

Zinc Intake	Mean	Standard Deviation	25th Percentile	50th Percentile	75th Percentile
Observed 3-day means	8.84	3.58	6.11	8.49	10.97
Adjusted intake	8.03	2.20	6.58	8.15	9.33

necessary to preserve the shape of the original distribution for analysis purposes while removing the within-person variance.

Table E-2 presents a comparison of the distribution of the observed subjects' 3-day means to the adjusted intake. The variance of the adjusted intake distribution is substantially less than the variance of the distribution of the observed 3-day means, as evidenced by the adjusted intake's lower standard deviation. In addition, the distance between the 25th and 75th percentiles of the adjusted intake distribution is closer to its mean than that of the observed 3-day mean.

If the Estimated Average Requirement (EAR) cut-point method is applied to the adjusted distribution to assess the prevalence of inadequate zinc intakes among this sample, an estimated 26 percent of women (12/46) appear to have inadequate intakes (12 of the 46 adjusted means were below the EAR for zinc for women of 6.8 mg/day). This is lower than the 28 percent prevalence of inadequacy that would be estimated from the unadjusted data.

Special Considerations

Two features of the NRC method deserve special note because they pose challenges to analysts wanting to use this approach. First is the requirement for normally distributed data, and the second is the handling of incomplete data.

Normality

As noted earlier, the NRC method hinges on having normally distributed intake data or being able to transform the observed data into a normal distribution. If nonnormal data are not transformed prior to adjustment, or if the applied transformation fails to correct for the nonnormality of the data, then assessments of the prevalence of inadequacy or excess using the adjusted distribution will be inaccurate. Some indication of the importance of this step comes from a closer look at the results of the adjustment procedure applied in the example presented above. Both the mean and the median of the adjusted distribution are slightly lower than the mean and median of the women's 3-day means (Table E-2), suggesting that the adjustment procedure has shifted the original distribution toward 0. This shift is a function of the transformation. Had the transformation more completely achieved the properties of a normal distribution, the observed mean and the adjusted mean would be equivalent.

It may be difficult, if not impossible, to normalize some observed nutrient intake distributions with simple power transformations. Observed distributions of vitamin A, in particular, are notorious for this problem (Aickin and Ritenbaugh, 1991; Beaton et al., 1983). In cases where the data fail to satisfy the assumptions of a normal distribution even when transformed, application of the NRC method and use of the resultant adjusted distribution for nutrient assessment is problematic (Beaton et al., 1997). Depending on the extent of the departure from normality, it may be preferable to not use the data for nutrient assessment. If assessments are conducted on data adjusted without fully satisfying the normality assumption, at minimum, the problem should be noted so that readers can interpret prevalence estimates with greater caution.

Handling Incomplete Data

The NRC method was originally developed for application to data sets with more than one day of intake data per subject. In describing the NRC method here, it has been assumed that an equal number of replicate observations are available for each member of the sample. If there are subjects missing one or more days of intake data, this can be factored into the calculation of $V_{between}$, reducing the denominator of that equation. Nonetheless, it is assumed that few subjects fall into this category.

In large dietary intake surveys it is increasingly common to collect two or more days of intake data on a subsample of the larger sample and use the understanding of within- and between-person variance derived from this subsample to adjust the intake data of the entire sample. (The ISU method [Nusser et al., 1996] is well suited to handling such data.) In surveys involving smaller samples, however, this practice is much less common. The application of estimates of within- and between-person variance from a subsample to the larger sample obviously presumes that the subsample is representative of the larger sample with respect to all characteristics that affect these variance estimates. If starting with a smaller sample, this representativeness may be more difficult to achieve through random sampling. With minor modifications to the NRC method outlined here it is possible to derive variance estimates from a subsample and apply this information to adjust the 1-day intake data for a larger sample. However, given the issue of representativeness, it is preferable to obtain two or more days of intake data on all subjects in a small sample and use all subjects' data in the adjustment procedure.

THE IOWA STATE UNIVERSITY METHOD

Working in conjunction with the U.S. Department of Agriculture, a group of statisticians at ISU developed a method to estimate usual intake distributions from large dietary surveys (Nusser et al., 1996). The method is implemented through a software package called SIDE (Software for Intake Distribution Estimation). It can be used to adjust observed intakes in large dietary surveys as long as two nonconsecutive or three consecutive days of intake data have been collected for a representative subsample of the group. For a full discussion of the ISU method of adjustment, see Guenther and colleagues (1997).

Based on the NRC method, the ISU approach includes a number of statistical enhancements (Guenther et al., 1997). Specifically, the ISU method is designed to transform the intakes for a nutrient to the standard normal distribution, applying procedures that go beyond the simple transformations that analysts can apply in the NRC method. The distribution of usual intakes is then estimated from this distribution of transformed intake values and the estimates are mapped back to the original scale through a bias-adjusted back transformation.

The procedures represent a major advance over the NRC method and a number of other more complicated adjustment procedures that have been proposed (Hoffmann et al., 2002). In addition, the ISU method is designed to take into account other factors such as day of week, time of year, and training or conditioning effects (apparent in patterns of reported intake in relation to the sequence of observations) that may exert systematic effects on the observed distribution of intakes. The ISU method can also account for correlation between observations on consecutive days and for heterogeneous within-person variances (e.g., in cases where the observed level of day-to-day variability in individuals' intakes is directly associated with their mean intake levels). While these refinements could be built into the NRC method, in its simplest form the method does not account for autocorrelation or other systematic effects on within-person variation.

Another particularly valuable feature of the ISU method is its ability to apply sample weighting factors, common in large population surveys, so that the adjusted distribution of intakes truly estimates the distribution of usual intakes in the target population, not just the sample. Thus the ISU method is well suited for use with large survey samples. In a recent evaluation of six different methods, Hoffmann and colleagues (2002) concluded that the ISU method had distinct advantages over the others. Most importantly, the method was applicable across a broad range of normally and nonnormally distributed intakes of food groups and nutrients.

Despite its strengths, however, the ISU method may not be as appropriate as the NRC method for use with small samples. The greater complexity of the ISU method requires a larger sample to ensure that the various steps in the adjustment procedure retain acceptable levels of reliability. A smaller sample can be used with the NRC method because the adjustment procedure is more simplistic (e.g., applying simpler methods of transformation and back-transformation and not accounting for heterogeneity of within-person variance).

OTHER CONSIDERATIONS IN THE APPLICATION OF ADJUSTMENT PROCEDURES

Defining Groups for Data Adjustment

Because nutrient requirements vary by life stage and gender group, assessments of nutrient adequacy are usually conducted separately for particular subgroups of the population. The statistical adjustment of intake data—whether done by the NRC or ISU method—should therefore also be conducted separately for each group for which the nutrient assessment will be conducted. If intake data have been collected across more than one life stage and gender group, it is not appropriate to combine subgroups for the purpose of adjustment and then later subdivide the adjusted data for separate analyses. Similarly, if the intended analysis of nutrient inadequacy is by stratum within a single life stage or gender group (e.g., the assessment of nutrient inadequacy for particular population subgroups defined by income or education levels), then the adjustment of intake data should be conducted separately for each stratum.

Adjusting Intake Variables Expressed as Ratios

To assess the macronutrient composition of diets and examine, for example, the proportion of energy derived from saturated fatty acids, it is necessary to examine the distribution of usual intakes for macronutrients expressed as ratios of total energy intake. The adjustment procedures described here can be applied to intakes expressed as nutrient:energy ratios or as nutrient:nutrient ratios. However, the ratio of interest should be computed for each day of intake data first; the observed intakes are then adjusted to estimate the distribution of usual intakes as ratios. For example, it is not appropriate to compute the adjusted distribution of energy and fat separately and then combine these distributions for analytic purposes.

Underlying Assumptions and Limitations of Adjustment Methods

One important difference in application of the two methods described here is that the ISU method of adjustment is typically applied to the distribution of intakes on day one of data collection, whereas the NRC method is applied to multiple-day means. In the design of large dietary surveys it is becoming increasingly common to collect a second day of intake data on only a subsample of the group. The ISU method is then applied to adjust the entire distribution of intakes on day one using the information about within-person variation that is gleaned from the subsample.

In the application of the NRC method to smaller data sets, typically comprising multiple days of intake data for each member of the sample, multiple-day means are used as the basis for adjustment with the underlying assumption that all days have equivalent validity. In data sets where a sequence effect is observed, with reported energy and nutrient intakes declining systematically across multiple days of data collection (Guenther et al., 1997), the adjustment of intakes to day-one data will result in a higher estimate of usual intake than an adjustment based on individuals' multiple-day means. If it can be assumed that intake on day one has been more accurately reported than on subsequent days, then clearly the adjustment to day-one data will yield a less biased estimate of the distribution of usual intakes. Because good methods to establish the validity of self-reported intakes on particular days of data collection are lacking, it is difficult to determine whether day-one data or multiple-day means are better estimates of true intake. Indeed, the answer may differ depending on the particular group under study and the conditions of data collection.

Neither the NRC nor the ISU method of adjustment is capable of addressing problems of systematic bias due to underreporting of intakes. The approaches must assume that individuals have reported their food intake without systematic bias—on day one, at least, for the ISU method, and across all days of data collection for the NRC method. If intakes have been underreported, the adjusted distribution of intakes will be biased by this underreporting.

Irrespective of the method of adjustment applied, it must also be assumed that reported food intakes have been correctly linked to a food composition database that accurately reflects the energy and nutrient content of the food. Systematic errors in the estimation of nutrient levels in foods consumed will bias the estimated distribution of usual intakes. In the case of nutrients for which food composition data are known to be incomplete, analysts must gauge the extent to which reported intakes will be biased. If intake cannot be estimated without substantial error, it is not appropriate to proceed with nutrient assessment.

Despite these limitations, the adjustment of observed distributions of intake for within-person variance to better estimate the distribution of usual intakes in a group represents a critical step in the assessment of nutrient adequacy or excess. In applying the steps in planning diets for groups, as described in this report, the focus is on planning for usual intakes. The assessments of nutrient adequacy and excess that are required to inform the planning process should be conducted on intake data that have been adjusted to provide the best possible estimate of the distribution of usual intakes in the group.

SAMPLE SAS PROGRAM FOR THE NRC METHOD

(Written by G.H. Beaton, University of Toronto, in December 1988 and modified in January 2002)

This program runs an ANOVA, estimates the partitioning of variance, and calculates the between-person, within-person, and total standard deviations (e.g., SDINTER, SDINTRA, and SDTOTAL, respectively) for the data set at hand with these estimates. The program then adjusts the observed distribution of mean intakes to remove remaining effects of within-person variation in intakes. The adjusted data can then be used as input data for the EAR cut-point or full probability assessment (IOM, 2000a). If the original data are transformed to better approximate a normal distribution, this program should be run on the transformed data and the final adjusted data back-transformed prior to the assessment of nutrient adequacy or excess. Note that the adjustments should be made independently for each stratification (e.g., males and females) and should be run on ratios after the ratio has been calculated.

```
************************************************************
**  NOTE: THIS PROGRAM, AS WRITTEN, ASSUMES THAT THE      **
**  INPUT DATA SET HAS ONE RECORD FOR EACH DAY OF         **
**  INTAKE.   IF MORE THAN ONE DAY OF INTAKE FOR EACH     **
**  SUBJECT APPEARS IN A SINGLE RECORD, THE DATA SET      **
**  WILL NEED TO BE REORGANIZED BEFORE THE PROGRAM        **
**  IS RUN.                                               **
************************************************************

PROC ANOVA DATA=YOURDATA OUTSTAT=ANOVSTAT;
CLASS SUBJID;
MODEL NUTRIENT=SUBJID;        *<< Change variable name to nutrient of
interest;
DATA PARTIT1;
SET ANOVSTAT;
MS = SS/DF;
MSERROR = MS; MSMODEL = MS;
DFERROR = DF; DFMODEL = DF;
IF _TYPE_ = ‘ERROR’ THEN MSMODEL = .;
IF _TYPE_ = ‘ANOVA’ THEN MSERROR = .;
IF _TYPE_ = ‘ERROR’ THEN DFMODEL = .;
IF _TYPE_ = ‘ANOVA’ THEN DFERROR = .;
KEEP MSMODEL DFMODEL MSERROR DFERROR;
PROC UNIVARIATE NOPRINT;
```

continued

```
VAR MSMODEL DFMODEL MSERROR DFERROR;
OUTPUT OUT=PARTIT2 MEAN = MSMODEL DFMODEL MSERROR
DFERROR;
DATA PARTIT3;
SET PARTIT2;
MEANREPL = (DFMODEL+DFERROR+1)/(DFMODEL+1);
ERRORDIF = MSMODEL - MSERROR;
IF ERRORDIF LT 0 THEN ERRORDIF = 0;
SDINTRA = MSERROR**0.5;
SDINTER = (ERRORDIF / MEANREPL)**0.5;
SDTOTAL = (SDINTER**2 +(SDINTRA**2/MEANREPL))**0.5;
INDEX=1;
KEEP SDINTER SDTOTAL INDEX;
PROC MEANS NOPRINT DATA=YOURDATA;
  VAR NUTRIENT; BY SUBJID;
   OUTPUT OUT=SUBJMEAN MEAN=SMEAN;
DATA SUBJMEAN; SET SUBJMEAN; INDEX=1;
PROC UNIVARIATE NOPRINT; VAR SMEAN;
OUTPUT OUT=MEANS  MEAN = GMEAN;
DATA MEANS; SET MEANS; INDEX=1;
DATA ADJUST;
MERGE SUBJMEAN PARTIT3 MEANS;
BY INDEX;
NRCADJ = GMEAN + (SMEAN - GMEAN) * SDINTER/SDTOTAL;
KEEP SUBJID NRCADJ;
RUN;

***************************************
** THIS IS NOW THE ADJUSTED           **
** DATA TO BE USED IN ANALYSIS        **
** NEED TO DO FOR EACH OF THE         **
** INTAKE VARIABLES IF THIS           **
** PROCEDURE IS TO BE EMPLOYED        **
***************************************

DATA FINAL; MERGE YOURDATA ADJUST; BY SUBJID;
PROC PRINT;
TITLE 'NUTRIENT DATA SHOWING INDIVIDUAL OBS, MEAN, NRC
ADJUSTED';
RUN;
```